Love Me, Love Me, Say (and Write!) that You Love Me: Enriching the WASABI Song Corpus with Lyrics Annotations

Michael Fell, Elena Cabrio, Elmahdi Korfed, Michel Buffa, Fabien Gandon

Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 2138–2147
Marseille, 11–16 May 2020

Love Me, Love Me, Say (and Write!) that You Love Me: Enriching the WASABI Song Corpus with Lyrics Annotations

Michael Fell, Elena Cabrio, Elmahdi Korfed, Michel Buffa and Fabien Gandon
Université Côte d'Azur, CNRS, Inria, I3S, France
{michael.fell, elena.cabrio, elmahdi.korfed, michel.buffa}@unice.fr, fabien.gandon@inria.fr

Abstract

We present the WASABI Song Corpus, a large corpus of songs enriched with metadata extracted from music databases on the Web, and resulting from the processing of song lyrics and from audio analysis. More specifically, given that lyrics encode an important part of the semantics of a song, we focus here on the description of the methods we proposed to extract relevant information from the lyrics, such as their structure segmentation, their topics, the explicitness of the lyrics content, the salient passages of a song and the emotions conveyed. The creation of the resource is still ongoing: so far, the corpus contains 1.73M songs with lyrics (1.41M unique lyrics) annotated at different levels with the output of the above mentioned methods. Such corpus labels and the provided methods can be exploited by music search engines and music professionals (e.g. journalists, radio presenters) to better handle large collections of lyrics, allowing an intelligent browsing, categorization and recommendation of songs. We provide the files of the current version of the WASABI Song Corpus, the models we have built on it as well as updates here: `https://github.com/micbuffa/WasabiDataset`.

Keywords: Corpus (Creation, Annotation, etc.), Information Extraction, Information Retrieval, Music and Song Lyrics

1. Introduction

Let's imagine the following scenario: following David Bowie's death, a journalist plans to prepare a radio show about the artist's musical career to acknowledge his qualities. To discuss the topic from different angles, she needs to have at her disposal the artist biographical information to know the history of his career, the song lyrics to know what he was singing about, his musical style, the emotions his songs were conveying, live recordings and interviews. Similarly, streaming professionals such as Deezer, Spotify, Pandora or Apple Music aim at enriching music listening with artists' information, to offer suggestions for listening to other songs/albums from the same or similar artists, or automatically determining the emotion felt when listening to a track to propose coherent playlists to the user. To support such scenarios, the need for rich and accurate musical knowledge bases and tools to explore and exploit this knowledge becomes evident.

In this paper, we present the WASABI Song Corpus, a large corpus of songs (2.10M songs, 1.73M with lyrics) enriched with metadata extracted from music databases on the Web, and resulting from the processing of song lyrics and from audio analysis. The corpus contains songs in 36 different languages, even if the vast majority are in English. As for the songs genres, the most common ones are Rock, Pop, Country and Hip Hop.

More specifically, while an overview of the goals of the WASABI project supporting the dataset creation and the description of a preliminary version of the dataset can be found in (Meseguer-Brocal et al., 2017), this paper focuses on the description of the methods we proposed to annotate relevant information in the song lyrics. Given that lyrics encode an important part of the semantics of a song, we propose to label the WASABI dataset lyrics with their structure segmentation, the explicitness of the lyrics content, the salient passages of a song, the addressed topics and the emotions conveyed.

An analysis of the correlations among the above mentioned annotation layers reveals interesting insights about the song corpus. For instance, we demonstrate the change in corpus annotations diachronically: we show that certain topics become more important over time and others are diminished. We also analyze such changes in explicit lyrics content and expressed emotion.

The paper is organized as follows. Section 2. introduces the WASABI Song Corpus and the metadata initially extracted from music databases on the Web. Section 3. describes the segmentation method we applied to decompose lyrics in their building blocks in the corpus. Section 4. explains the method used to summarize song lyrics, leveraging their structural properties. Section 5. reports on the annotations resulting from the explicit content classifier, while Section 6. describes how information on the emotions are extracted from the lyrics. Section 7. describes the topic modeling algorithm to label each lyrics with the top 10 words, while Section 8. examines the changes in the annotations over the course of time. Section 9. reports on similar existing resources, while Section 10. concludes the paper.

2. The WASABI Song Corpus

In the context of the WASABI research project[1] that started in 2017, a two million song database has been built, with metadata on 77k artists, 208k albums, and 2.10M songs (Meseguer-Brocal et al., 2017). The metadata has been *i)* aggregated, merged and curated from different data sources on the Web, and *ii)* enriched by pre-computed or on-demand analyses of the lyrics and audio data.

We have performed various levels of analysis, and interactive Web Audio applications have been built on top of the output. For example, the TimeSide analysis and annotation framework have been linked (Fillon et al., 2014) to make on-demand audio analysis possible. In connection

[1] `http://wasabihome.i3s.unice.fr/`

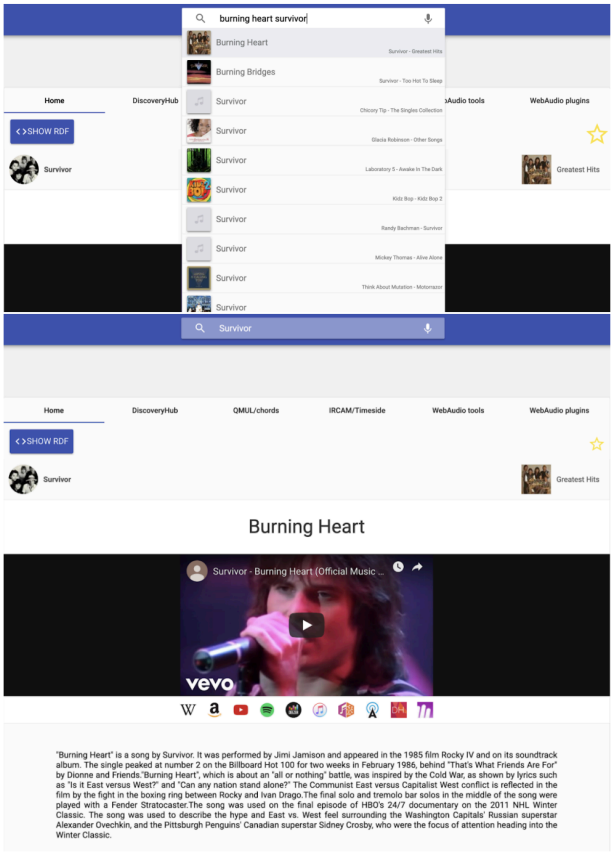


Figure 1: The WASABI Interactive Navigator.

with the FAST project[2], an offline chord analysis of 442k songs has been performed, and both an online enhanced audio player (Pauwels and Sandler, 2019) and chord search engine (Pauwels et al., 2018) have been built around it. A rich set of Web Audio applications and plugins has been proposed (Buffa and Lebrun, 2017a; Buffa and Lebrun, 2017b; Buffa et al., 2018), that allow, for example, songs to be played along with sounds similar to those used by artists. All these metadata, computational analyses and Web Audio applications have now been gathered in one easy-to-use web interface, the WASABI Interactive Navigator[3], illustrated[4] in Figure 1.

We have started building the WASABI Song Corpus by collecting for each artist the complete discography, band members with their instruments, time line, equipment they use, and so on. For each song we collected its lyrics from LyricWiki[5], the synchronized lyrics when available[6], the DBpedia abstracts and the categories the song belongs to, e.g. genre, label, writer, release date, awards, producers, artist and band members, the stereo audio track from Deezer, the unmixed audio tracks of the song, its ISRC, bpm and duration.

We matched the song ids from the WASABI Song Corpus with the ids from MusicBrainz, iTunes, Discogs, Spotify, Amazon, AllMusic, GoHear, YouTube. Figure 2 illustrates[7] all the data sources we have used to create the WASABI Song Corpus. We have also aligned the WASABI Song Corpus with the publicly available LastFM dataset[8], resulting in 327k tracks in our corpus having a LastFM id. As of today, the corpus contains 1.73M songs with lyrics (1.41M unique lyrics). 73k songs have at least an abstract on DBpedia, and 11k have been identified as "classic songs" (they have been number one, or got a Grammy award, or have lots of cover versions). About 2k songs have a multi-track audio version, and on-demand source separation using open-unmix (Stöter et al., 2019) or Spleeter (Hennequin et al., 2019) is provided as a TimeSide plugin. Several Natural Language Processing methods have been applied to the lyrics of the songs included in the WASABI Song Corpus, as well as various analyses of the extracted information have been carried out. After providing some statistics on the WASABI corpus, the rest of the article describes the different annotations we added to the lyrics of the songs in the dataset. Based on the research we have conducted, the following lyrics annotations are added: lyrical structure (Section 3.), summarization (Section 4.), explicit lyrics (Section 5.), emotion in lyrics (Section 6.) and topics in lyrics (Section 7.).

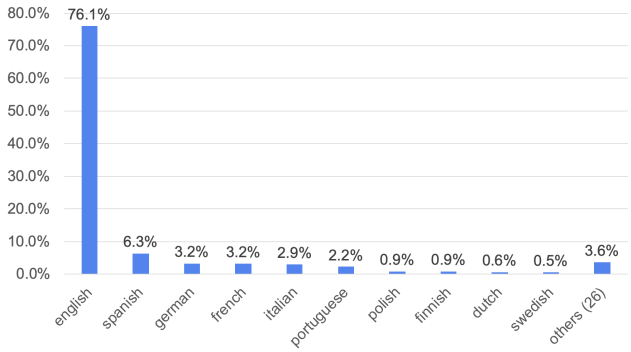
Figure 2: The datasources connected to the WASABI Song Corpus.

2.1. Statistics on the WASABI Song Corpus

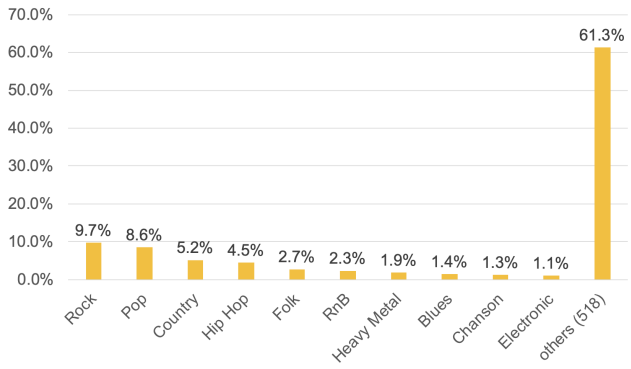
This section summarizes key statistics on the corpus, such as the language and genre distributions, the songs coverage in terms of publication years, and then gives the technical details on its accessibility.

Language Distribution Figure 3a shows the distribution of the ten most frequent languages in our corpus.[9] In to-

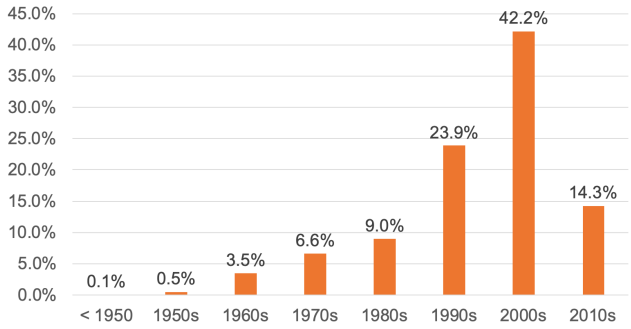
[2] http://www.semanticaudio.ac.uk
[3] http://wasabi.i3s.unice.fr/
[4] Illustration taken from (Buffa et al., 2019a).
[5] http://lyrics.wikia.com/
[6] from http://usdb.animux.de/
[7] Illustration taken from (Buffa et al., 2019b).
[8] http://millionsongdataset.com/lastfm/
[9] Based on language detection performed on the lyrics.



(a) Language distribution (100% = 1.73M)



(b) Genre distribution (100% = 1.06M)



(c) Decade of publication distribution (100% = 1.70M)

Figure 3: Statistics on the WASABI Song Corpus

tal, the corpus contains songs of 36 different languages. The vast majority (76.1%) is English, followed by Spanish (6.3%) and by four languages in the 2-3% range (German, French, Italian, Portugese). On the bottom end, Swahili and Latin amount to 0.1% (around 2k songs) each.

Genre Distribution In Figure 3b we depict the distribution of the ten most frequent genres in the corpus.[10] In total, 1.06M of the titles are tagged with a genre. It should be noted that the genres are very sparse with a total of 528 different ones. This high number is partially due to many subgenres such as Alternative Rock, Indie Rock, Pop Rock, etc. which we omitted in Figure 3b for clarity. The most common genres are Rock (9.7%), Pop (8.6%), Country (5.2%), Hip Hop (4.5%) and Folk (2.7%).

Publication Year Figure 3c shows the number of songs published in our corpus, by decade.[11] We find that over 50% of all songs in the WASABI Song Corpus are from the 2000s or later and only around 10% are from the seventies or earlier.

Accessibility of the WASABI Song Corpus The WASABI Interactive Navigator relies on multiple database engines: it runs on a MongoDB server altogether with an indexation by Elasticsearch and also on a Virtuoso triple store as a RDF graph database. It comes with a REST API[12] and an upcoming SPARQL endpoint. All the database metadata is publicly available[13] under a CC licence through the WASABI Interactive Navigator as well as programmatically through the WASABI REST API.

We provide the files of the current version of the WASABI Song Corpus, the models we have built on it as well as updates here: `https://github.com/micbuffa/WasabiDataset`.

3. Lyrics Structure Annotations

Generally speaking, lyrics structure segmentation consists of two stages: text segmentation to divide lyrics into segments, and semantic labelling to label each segment with a structure type (e.g. Intro, Verse, Chorus).

In (Fell et al., 2018) we proposed a method to segment lyrics based on their repetitive structure in the form of a self-similarity matrix (SSM). Figure 4 shows a line-based SSM for the song text written on top of it[14]. The lyrics consists of seven segments and shows the typical repetitive structure of a Pop song. The main diagonal is trivial, since each line is maximally similar to itself. Notice further the additional diagonal stripes in segments 2, 4 and 7; this indicates a repeated part, typically the chorus. Based on the simple idea that eyeballing an SSM will reveal (parts of) a song's structure, we proposed a Convolutional Neural Network architecture that successfully learned to predict segment borders in the lyrics when "looking at" their SSM. Table 1 shows the genre-wise results we obtained using our proposed architecture. One important insight was that more repetitive lyrics as often found in genres such as Country and Punk Rock are much easier to segment than lyrics in Rap or Hip Hop which often do not even contain a chorus.

In the WASABI Interactive Navigator, the line-based SSM of a song text can be visualized. It is toggled by clicking on the violet-blue square on top of the song text. For a subset of songs the color opacity indicates how repetitive and representative a segment is, based on the fitness metric that we proposed in (Fell et al., 2019b). Note how in Figure 4,

[10] We take the genre of the album as ground truth since song-wise genres are much rarer.

[11] We take the album publication date as proxy since song-wise labels are too sparse.

[12] `https://wasabi.i3s.unice.fr/apidoc/`

[13] There is no public access to copyrighted data such as lyrics and full length audio files. Instructions on how to obtain lyrics are nevertheless provided and audio extracts of 30s length are available for nearly all songs.

[14] `https://wasabi.i3s.unice.fr/#/search/artist/Britney%20Spears/album/In%20The%20Zone/song/Everytime`

Genre	*P*	*R*	F_1
Rock	73.8	57.7	64.8
Hip Hop	71.7	43.6	54.2
Pop	73.1	61.5	66.6
RnB	71.8	60.3	65.6
Alternative Rock	76.8	60.9	67.9
Country	74.5	66.4	**70.2**
Hard Rock	76.2	61.4	67.7
Pop Rock	73.3	59.6	65.8
Indie Rock	80.6	55.5	65.6
Heavy Metal	79.1	52.1	63.0
Southern Hip Hop	73.6	34.8	47.0
Punk Rock	80.7	63.2	**70.9**
Alternative Metal	77.3	61.3	68.5
Pop Punk	77.3	68.7	**72.7**
Gangsta Rap	73.6	35.2	47.7
Soul	70.9	57.0	63.0

Table 1: Lyrics segmentation performances across musical genres in terms of Precision (*P*), Recall (*R*) and F_1 in %. Underlined are the performances on genres with less repetitive text. Genres with highly repetitive structure are in bold.

Figure 4: Structure of the lyrics of "Everytime" by Britney Spears as displayed in the WASABI Interactive Navigator.

the segments 2, 4 and 7 are shaded more darkly than the surrounding ones. As highly fit (opaque) segments often coincide with a chorus, this is a first approximation of chorus detection. Given the variability in the set of structure types provided in the literature according to different genres (Tagg, 1982; Brackett, 1995), rare attempts have been made in the literature to achieve a more complete semantic labelling, labelling the lyrics segments as Intro, Verse, Bridge, Chorus etc.

For each song text we provide an SSM based on a normalized character-based edit distance[15] on two levels of granularity to enable other researchers to work with these structural representations: line-wise similarity and segment-wise similarity.

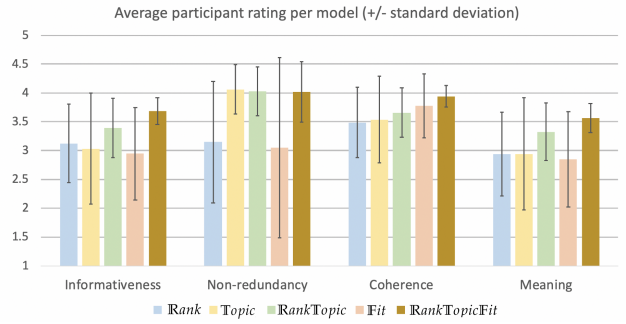


Figure 5: Human ratings per summarization model (five point Likert scale). Models are $\mathbb{R}ank$: graph-based, $\mathbb{T}opic$: topic-based, $\mathbb{F}it$: thumbnail-based, and model combinations.

4. Lyrics Summary

Given the repeating forms, peculiar structure and other unique characteristics of song lyrics, in (Fell et al., 2019b) we introduced a method for extractive summarization of lyrics that takes advantage of these additional elements to more accurately identify relevant information in song lyrics. More specifically, it relies on the intimate relationship between the audio and the lyrics. The so-called audio thumbnails, snippets of usually 30 seconds of music, are a popular means to summarize a track in the audio community. The intuition is the more repeated and the longer a part, the better it represents the song. We transferred an audio thumbnailing approach to our domain of lyrics and showed that adding the thumbnail improves summary quality. We evaluated our method on 50k lyrics belonging to the top 10 genres of the WASABI Song Corpus and according to qualitative criteria such as *Informativeness* and *Coherence*. Figure 5 shows our results for different summarization models. Our model $\mathbb{R}ank\mathbb{T}opic\mathbb{F}it$, which combines graph-based, topic-based and thumbnail-based summarization, outperforms all other summarizers. We further find that the genres RnB and Country are highly overrepresented in the lyrics sample with respect to the full WASABI Song Corpus, indicating that songs belonging to these genres are more likely to contain a chorus. Finally, Figure 6 shows an example summary of four lines length obtained with our proposed $\mathbb{R}ank\mathbb{T}opic\mathbb{F}it$ method. It is toggled in the WASABI Interactive Navigator by clicking on the green square on top of the song text.

The four-line summaries of 50k English used in our experiments are freely available within the WASABI Song Corpus; the Python code of the applied summarization methods is also available[16].

[15] In our segmentation experiments we found this simple metric to outperform more complex metrics that take into account the phonetics or the syntax.

[16] https://github.com/TuringTrain/lyrics_thumbnailing

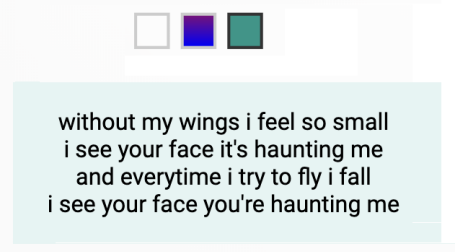


Figure 6: Summary of the lyrics of "Everytime" by Britney Spears as displayed in the WASABI Interactive Navigator.

5. Explicit Language in Lyrics

On audio recordings, the Parental Advisory Label is placed in recognition of profanity and to warn parents of material potentially unsuitable for children. Nowadays, such labelling is carried out mainly manually on voluntary basis, with the drawbacks of being time consuming and therefore costly, error prone and partly a subjective task. In (Fell et al., 2019a) we have tackled the task of automated explicit lyrics detection, based on the songs carrying such a label. We compared automated methods ranging from dictionary-based lookup to state-of-the-art deep neural networks to automatically detect explicit contents in English lyrics. More specifically, the dictionary-based methods rely on a swear word dictionary D_n which is automatically created from example explicit and clean lyrics. Then, we use D_n to predict the class of an unseen song text in one of two ways: (i) the *Dictionary Lookup* simply checks if a song text contains words from D_n. (ii) the *Dictionary Regression* uses BOW made from D_n as the feature set of a logistic regression classifier. In the *Tf-idf BOW Regression* the BOW is expanded to the whole vocabulary of a training sample instead of only the explicit terms. Furthermore, the model *TDS Deconvolution* is a deconvolutional neural network (Vanni et al., 2018) that estimates the importance of each word of the input for the classifier decision. In our experiments, we worked with 179k lyrics that carry gold labels provided by Deezer (17k tagged as explicit) and obtained the results shown in Figure 2. We found the very simple *Dictionary Lookup* method to perform on par with much more complex models such as the *BERT Language Model* (Devlin et al., 2018) as a text classifier. Our analysis revealed that some genres are highly overrepresented among the explicit lyrics. Inspecting the automatically induced explicit words dictionary reflects that genre bias. The dictionary of 32 terms used for the dictionary lookup method consists of around 50% of terms specific to the Rap genre, such as glock, gat, clip (gun-related), thug, beef, gangsta, pimp, blunt (crime and drugs). Finally, the terms holla, homie, and rapper are obviously no swear words, but highly correlated with explicit content lyrics.

Our corpus contains 52k tracks labelled as explicit and 663k clean (not explicit) tracks[17]. We have trained a classifier (77.3% f-score on test set) on the 438k English lyrics which are labelled and classified the remaining 455k previously untagged English tracks. We provide both the predicted labels in the WASABI Song Corpus and the trained classifier to apply it to unseen text.

Model	*P*	*R*	F_1
Majority Class	45.0	50.0	47.4
Dictionary Lookup	78.3	76.4	77.3
Dictionary Regression	76.2	81.5	78.5
Tf-idf BOW Regression	75.6	81.2	78.0
TDS Deconvolution	81.2	78.2	79.6
BERT Language Model	84.4	73.7	77.7

Table 2: Performance comparison of our different models. Precision (*P*), Recall (*R*) and f-score (F_1) in %.

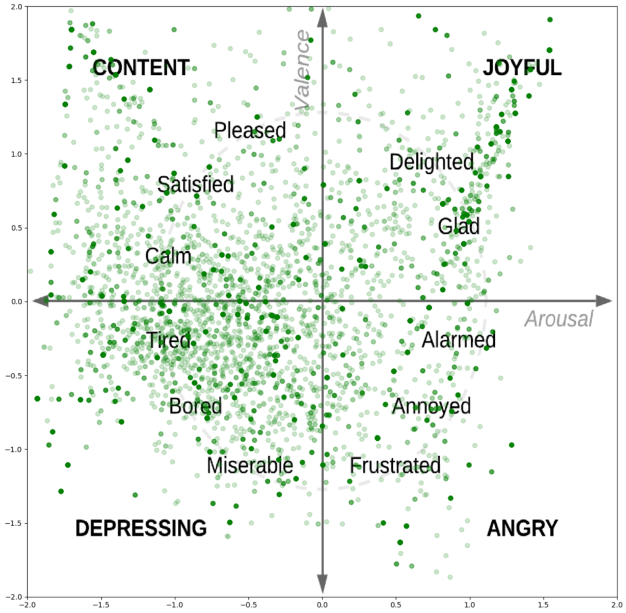


Figure 7: Emotion distribution in the corpus in the valence-arousal plane.

6. Emotional Description

In sentiment analysis the task is to predict if a text has a positive or a negative emotional valence. In the recent years, a transition from detecting sentiment (positive vs. negative valence) to more complex formulations of emotion detection (e.g. joy, fear, surprise) (Mohammad et al., 2018) has become more visible; even tackling the problem of emotion in context (Chatterjee et al., 2019). One family of emotion detection approaches is based on the valence-arousal model of emotion (Russell, 1980), locating every emotion in a two-dimensional plane based on its valence (positive vs. negative) and arousal (aroused vs. calm).[18] Figure 7 is an illustration of the valence-arousal model of Russell and shows exemplary where several emotions such as joyful, angry or calm are located in the plane. Manually labelling texts with multi-dimensional emotion descriptions is an inherently hard task. Therefore, researchers have resorted to distant supervision, obtaining gold labels from social tags from lastfm. These approaches (Hu et al., 2009a; Çano and

[17] Labels provided by Deezer. Furthermore, 625k songs have a different status such as unknown or censored version.

[18] Sometimes, a third dimension of dominance is part of the model.

Morisio, May 2017) define a list of social tags that are related to emotion, then project them into the valence-arousal space using an emotion lexicon (Warriner et al., 2013; Mohammad, 2018).

Recently, Deezer made valence-arousal annotations for 18,000 English tracks available[19] they have derived by the aforementioned method (Delbouys et al., 2018). We aligned the valence-arousal annotations of Deezer to our songs. In Figure 7 the green dots visualize the emotion distribution of these songs.[20] Based on their annotations, we train an emotion regression model using BERT, with an evaluated 0.44/0.43 Pearson correlation/Spearman correlation for valence and 0.33/0.31 for arousal on the test set.

We integrated Deezer's labels into our corpus and also provide the valence-arousal predictions for the 1.73M tracks with lyrics. We also provide the last.fm social tags (276k) and emotion tags (87k entries) to facilitate researchers to build variants of emotion recognition models.

7. Topic Modelling

We built a topic model on the lyrics of our corpus using Latent Dirichlet Allocation (LDA) (Blei et al., 2003). We determined the hyperparameters α, η and the topic count such that the coherence was maximized on a subset of 200k lyrics. We then trained a topic model of 60 topics on the unique English lyrics (1.05M).

We have manually labelled a number of more recognizable topics. Figures 9-13 illustrate these topics with word clouds[21] of the most characteristic words per topic. For instance, the topic Money contains words of both the field of earning money (job, work, boss, sweat) as well as spending it (pay, buy). The topic Family is both about the people of the family (mother, daughter, wife) and the land (sea, valley, tree).

We provide the topic distribution of our LDA topic model for each song and make available the trained topic model to enable its application to unseen lyrics.

8. Diachronic Corpus Analysis

We examine the changes in the annotations over the course of time by grouping the corpus into decades of songs according to the distribution shown in Figure 3c.

Changes in Topics The importance of certain topics has changed over the decades, as depicted in Figure 14a. Some topics have become more important, others have declined, or stayed relatively the same. We define the importance of a topic for a decade of songs as follows: first, the LDA topic model trained on the full corpus gives the probability of the topic for each song separately. We then average these songwise probabilities over all songs of the decade. For each of the cases of growing, diminishing and constant importance, we display two topics. The topics War and Death have appreciated in importance over time. This is partially caused by the rise of Heavy Metal in the beginning of the 1970s, as the vocabulary of the Death topic is very typical for the

Figure 8: Topic War

Figure 9: Topic Death

Figure 10: Topic Love

Figure 11: Topic Family

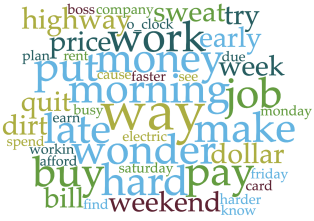


Figure 12: Topic Money

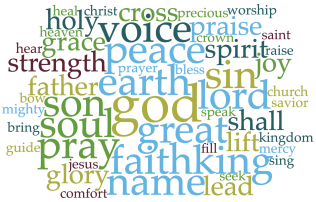


Figure 13: Topic Religion

genre (see for instance the "Metal top 100 words" in (Fell and Sporleder, 2014)). We measure a decline in the importance of the topics Love and Family. The topics Money and Religion seem to be evergreens as their importance stayed rather constant over time.

Changes in Explicitness We find that newer songs are more likely being tagged as having explicit content lyrics. Figure 14b shows our estimates of explicitness per decade, the ratio of songs in the decade tagged as explicit to all songs of the decade. Note that the Parental Advisory Label was first distributed in 1985 and many older songs may not have been labelled retroactively. The depicted evolution of explicitness may therefore overestimate the "true explicitness" of newer music and underestimate it for music before 1985.

Changes in Emotion We estimate the emotion of songs in a decade as the average valence and arousal of songs of that decade. We find songs to decrease both in valence and arousal over time. This decrease in positivity (valence) is in line with the diminishment of positively connotated topics such as Love and Family and the appreciation of topics with a more negative connotation such as War and Death.

9. Related Work

This section describes available songs and lyrics databases, and summarizes existing work on lyrics processing.

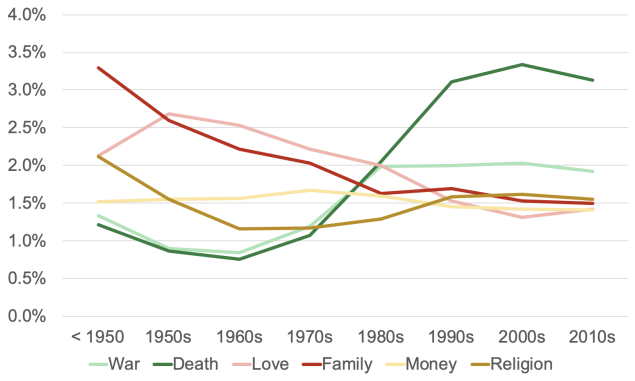
Songs and Lyrics Databases. The Million Song Dataset (MSD) project[22] (Bertin-Mahieux et al., 2011) is a collection of audio features and metadata for a million contempo-

[19] https://github.com/deezer/deezer_mood_detection_dataset

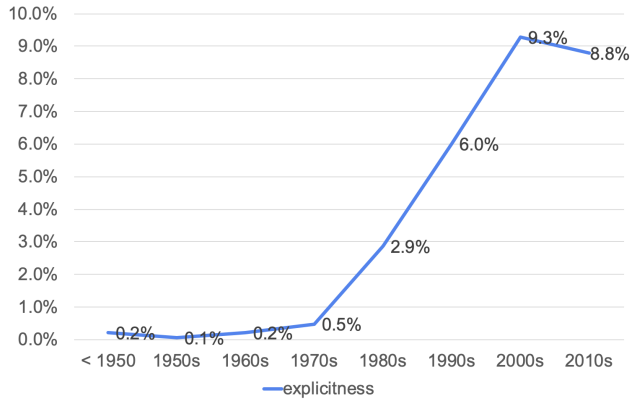
[20] Depiction without scatterplot taken from (Parisi et al., 2019)

[21] made with https://www.wortwolken.com/

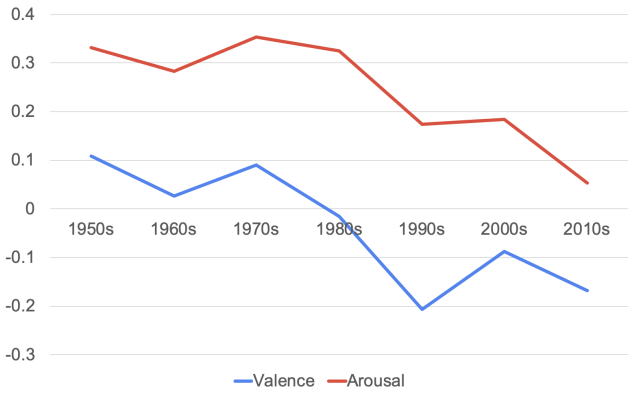
[22] http://millionsongdataset.com



(a) Evolution of topic importance



(b) Evolution of explicit content lyrics



(c) Evolution of emotion

Figure 14: Evolution of different annotations during the decades

rary popular music tracks. Such dataset shares some similarities with WASABI with respect to metadata extracted from Web resources (as artist names, tags, years) and audio features, even if at a smaller scale. Given that it mainly focuses on audio data, a complementary dataset providing lyrics of the Million Song dataset was released, called musiXmatch dataset[23]. It consists in a collection of song lyrics in bag-of-words (plus stemmed words), associated with MSD tracks. However, no other processing of the lyrics is done, as is the case in our work.

MusicWeb and its successor MusicLynx (Allik et al., 2018) link music artists within a Web-based application for discovering connections between them and provides a browsing experience using extra-musical relations. The project shares some ideas with WASABI, but works on the artist level, and does not perform analyses on the audio and lyrics content itself. It reuses, for example, MIR metadata from AcousticBrainz.

The WASABI project has been built on a broader scope than these projects and mixes a wider set of metadata, including ones from audio and natural language processing of lyrics. In addition, as presented in this paper, it comes with a large set of Web Audio enhanced applications (multitrack player, online virtual instruments and effect, on-demand audio processing, audio player based on extracted, synchronized chords, etc.)

Companies such as Spotify, GraceNote, Pandora, or Apple Music have sophisticated private knowledge bases of songs and lyrics to feed their search and recommendation algorithms, but such data are not available (and mainly rely on audio features).

Lyrics Segmentation. Only a few papers in the literature have focused on the automated detection of the structure of lyrics. (Watanabe et al., 2016) propose the task to automatically identify segment boundaries in lyrics and train a logistic regression model for the task with the repeated pattern and textual features. (Mahedero et al., 2005) report experiments on the use of standard NLP tools for the analysis of music lyrics. Among the tasks they address, for structure extraction they focus on a small sample of lyrics having a clearly recognizable structure (which is not always the case) divided into segments. More recently, (Baratè et al., 2013) describe a semantics-driven approach to the automatic segmentation of song lyrics, and mainly focus on pop/rock music. Their goal is not to label a set of lines in a given way (e.g. verse, chorus), but rather identifying recurrent as well as non-recurrent groups of lines. They propose a rule-based method to estimate such structure labels of segmented lyrics.

Explicit Content Detection. (Bergelid, 2018) consider a dataset of English lyrics to which they apply classical machine learning algorithms. The explicit labels are obtained from Soundtrack Your Brand [24]. They also experiment with adding lyrics metadata to the feature set, such as the artist name, the release year, the music energy level, and the valence/positiveness of a song. (Chin et al., 2018) apply explicit lyrics detection to Korean song texts. They also use tf-idf weighted BOW as lyrics representation and aggregate multiple decision trees via boosting and bagging to classify the lyrics for explicit content. More recently, (Kim and Mun, 2019) proposed a neural network method to create explicit words dictionaries automatically by weighting a vocabulary according to all words' frequencies in the explicit class vs. the clean class, accordingly. They work with a corpus of Korean lyrics.

[23] http://millionsongdataset.com/musixmatch/

[24] https://www.soundtrackyourbrand.com

Emotion Recognition Recently, (Delbouys et al., 2018) address the task of multimodal music mood prediction based on the audio signal and the lyrics of a track. They propose a new model based on deep learning outperforming traditional feature engineering based approaches. Performances are evaluated on their published dataset with associated valence and arousal values which we introduced in Section 6.

(Xia et al., 2008) model song texts in a low-dimensional vector space as bags of concepts, the "emotional units"; those are combinations of emotions, modifiers and negations. (Yang and Lee, 2009) leverage the music's emotion annotations from Allmusic which they map to a lower dimensional psychological model of emotion. They train a lyrics emotion classifier and show by qualitative interpretation of an ablated model (decision tree) that the deciding features leading to the classes are intuitively plausible. (Hu et al., 2009b) aim to detect emotions in song texts based on Russell's model of mood; rendering emotions continuously in the two dimensions of arousal and valence (positive/negative). They analyze each sentence as bag of "emotional units"; they reweight sentences' emotions by both adverbial modifiers and tense and even consider progressing and adversarial valence in consecutive sentences. Additionally, singing speed is taken into account. With the fully weighted sentences, they perform clustering in the 2D plane of valence and arousal. Although the method is unsupervised at runtime, there are many parameters tuned manually by the authors in this work.

(Mihalcea and Strapparava, 2012) render emotion detection as a multi-label classification problem, songs express intensities of six different basic emotions: anger, disgust, fear, joy, sadness, surprise. Their corpus (100 song texts) has time-aligned lyrics with information on musical key and note progression. Using Mechanical Turk they each line of song text is annotated with the six emotions. For emotion classification, they use bags of words and concepts, as musical features key and notes. Their classification results using both modalities, textual and audio features, are significantly improved compared to a single modality.

Topic Modelling Among the works addressing this task for song lyrics, (Mahedero et al., 2005) define five ad hoc topics (Love, Violent, Antiwar, Christian, Drugs) into which they classify their corpus of 500 song texts using supervision. Related, (Fell, 2014) also use supervision to find bags of genre-specific n-grams. Employing the view from the literature that BOWs define topics, the genre-specific terms can be seen as mixtures of genre-specific topics.

(Logan et al., 2004) apply the unsupervised topic model Probabilistic LSA to their ca. 40k song texts. They learn latent topics for both the lyrics corpus as well as a NYT newspaper corpus (for control) and show that the domain-specific topics slightly improve the performance in their MIR task. While their MIR task performs highly better when using acoustic features, they discover that both methods err differently. (Kleedorfer et al., 2008) apply Non-negative Matrix Factorization (NMF) to ca. 60k song texts and cluster them into 60 topics. They show the so discovered topics to be intrinsically meaningful.

(Sterckx, 2014) have worked on topic modelling of a large-scale lyrics corpus of 1M songs. They build models using Latent Dirichlet allocation with topic counts between 60 and 240 and show that the 60 topics model gives a good trade-off between topic coverage and topic redundancy. Since popular topic models such as LDA represent topics as weighted bags of words, these topics are not immediately interpretable. This gives rise to the need of an automatic labelling of topics with smaller labels. A recent approach (Bhatia et al., 2016) relates the topical BOWs with titles of Wikipedia articles in a two step procedure: first, candidates are generated, then ranked.

10. Conclusion

In this paper we have described the WASABI dataset of songs, focusing in particular on the lyrics annotations resulting from the applications of the methods we proposed to extract relevant information from the lyrics. So far, lyrics annotations concern their structure segmentation, their topic, the explicitness of the lyrics content, the summary of a song and the emotions conveyed. Some of those annotation layers are provided for all the 1.73M songs included in the WASABI corpus, while some others apply to subsets of the corpus, due to various constraints described in the paper. Table 3 summarizes the most relevant annotations in our corpus.

Annotation	*Labels*	*Description*
Lyrics	1.73M	segments of lines of text
Languages	1.73M	36 different ones
Genre	1.06M	528 different ones
Last FM id	326k	UID
Structure	1.73M	SSM $\in \mathbb{R}^{n \times n}$ (n: length)
$\mathbb{S}$ocial tags	276k	$\mathbb{S} = \{$rock, joyful, 90s, ...$\}$
$\mathbb{E}$motion tags	87k	$\mathbb{E} \subset \mathbb{S} = \{$joyful, tragic, ...$\}$
Explicitness	715k	True (52k), False (663k)
Explicitness ♣	455k	True (85k), False (370k)
Summary♣	50k	four lines of song text
Emotion	16k	(valence, arousal) $\in \mathbb{R}^2$
Emotion♣	1.73M	(valence, arousal) $\in \mathbb{R}^2$
Topics♣	1.05M	Prob. distrib. $\in \mathbb{R}^{60}$
Total tracks	2.10M	diverse metadata

Table 3: Most relevant song-wise annotations in the WASABI Song Corpus. Annotations with ♣ are predictions of our models.

As the creation of the resource is still ongoing, we plan to integrate an improved emotional description in future work. In (Atherton and Kaneshiro, 2016) the authors have studied how song writers influence each other. We aim to learn a model that detects the border between heavy influence and plagiarism.

Acknowledgement

This work is partly funded by the French Research National Agency (ANR) under the WASABI project (contract ANR-16-CE23-0017-01) and by the EU Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant agreement No. 690974 (MIREL).

11. Bibliographical References

Allik, A., Thalmann, F., and Sandler, M. (2018). MusicLynx: Exploring music through artist similarity graphs. In *Companion Proc. (Dev. Track) The Web Conf. (WWW 2018)*.

Atherton, J. and Kaneshiro, B. (2016). I said it first: Topological analysis of lyrical influence networks. In *ISMIR*, pages 654–660.

Baratè, A., Ludovico, L. A., and Santucci, E. (2013). A semantics-driven approach to lyrics segmentation. In *2013 8th International Workshop on Semantic and Social Media Adaptation and Personalization*, pages 73–79, Dec.

Bergelid, L. (2018). Classification of explicit music content using lyrics and music metadata.

Bhatia, S., Lau, J. H., and Baldwin, T. (2016). Automatic labelling of topics with neural embeddings. *arXiv preprint arXiv:1612.05340*.

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022.

Brackett, D. (1995). *Interpreting Popular Music*. Cambridge University Press.

Buffa, M. and Lebrun, J. (2017a). Real time tube guitar amplifier simulation using webaudio. In *Proc. 3rd Web Audio Conference (WAC 2017)*.

Buffa, M. and Lebrun, J. (2017b). Web audio guitar tube amplifier vs native simulations. In *Proc. 3rd Web Audio Conf. (WAC 2017)*.

Buffa, M., Lebrun, J., Kleimola, J., Letz, S., et al. (2018). Towards an open web audio plugin standard. In *Companion Proceedings of the The Web Conference 2018*, pages 759–766. International World Wide Web Conferences Steering Committee.

Buffa, M., Lebrun, J., Pauwels, J., and Pellerin, G. (2019a). A 2 Million Commercial Song Interactive Navigator. In *WAC 2019 - 5th WebAudio Conference 2019*, Trondheim, Norway, December.

Buffa, M., Lebrun, J., Pellerin, G., and Letz, S. (2019b). Webaudio plugins in daws and for live performance. In *14th International Symposium on Computer Music Multidisciplinary Research (CMMR'19)*.

Çano, E. and Morisio, M. (May 2017). Music mood dataset creation based on last.fm tags. In *2017 International Conference on Artificial Intelligence and Applications, Vienna Austria*.

Chatterjee, A., Narahari, K. N., Joshi, M., and Agrawal, P. (2019). Semeval-2019 task 3: Emocontext contextual emotion detection in text. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 39–48.

Chin, H., Kim, J., Kim, Y., Shin, J., and Yi, M. Y. (2018). Explicit content detection in music lyrics using machine learning. In *2018 IEEE International Conference on Big Data and Smart Computing (BigComp)*, pages 517–521. IEEE.

Delbouys, R., Hennequin, R., Piccoli, F., Royo-Letelier, J., and Moussallam, M. (2018). Music mood detection based on audio and lyrics with deep neural net. *arXiv preprint arXiv:1809.07276*.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Fell, M. and Sporleder, C. (2014). Lyrics-based analysis and classification of music. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 620–631.

Fell, M., Nechaev, Y., Cabrio, E., and Gandon, F. (2018). Lyrics Segmentation: Textual Macrostructure Detection using Convolutions. In *Conference on Computational Linguistics (COLING)*, pages 2044–2054, Santa Fe, New Mexico, United States, August.

Fell, M., Cabrio, E., Corazza, M., and Gandon, F. (2019a). Comparing Automated Methods to Detect Explicit Content in Song Lyrics. In *RANLP 2019 - Recent Advances in Natural Language Processing*, Varna, Bulgaria, September.

Fell, M., Cabrio, E., Gandon, F., and Giboin, A. (2019b). Song lyrics summarization inspired by audio thumbnailing. In *RANLP 2019 - Recent Advances in Natural Language Processing (RANLP)*, Varna, Bulgaria, September.

Fell, M. (2014). Lyrics classification. In *Master's thesis, Saarland University, Germany, 2014.*, 01.

Fillon, T., Simonnot, J., Mifune, M.-F., Khoury, S., Pellerin, G., and Le Coz, M. (2014). Telemeta: An open-source web framework for ethnomusicological audio archives management and automatic analysis. In *Proceedings of the 1st International Workshop on Digital Libraries for Musicology*, pages 1–8. ACM.

Hennequin, R., Khlif, A., Voituret, F., and Moussallam, M. (2019). Spleeter: A fast and state-of-the art music source separation tool with pre-trained models. Late-Breaking/Demo ISMIR 2019, November. Deezer Research.

Hu, X., Downie, J. S., and Ehmann, A. F. (2009a). Lyric text mining in music mood classification. *American music*, 183(5,049):2–209.

Hu, Y., Chen, X., and Yang, D. (2009b). Lyric-based song emotion detection with affective lexicon and fuzzy clustering method. In *ISMIR*.

Kim, J. and Mun, Y. Y. (2019). A hybrid modeling approach for an automated lyrics-rating system for adolescents. In *European Conference on Information Retrieval*, pages 779–786. Springer.

Kleedorfer, F., Knees, P., and Pohle, T. (2008). Oh oh oh whoah! towards automatic topic detection in song lyrics. In *ISMIR*.

Logan, B., Kositsky, A., and Moreno, P. (2004). Semantic analysis of song lyrics. In *2004 IEEE International Conference on Multimedia and Expo (ICME) (IEEE Cat. No.04TH8763)*, volume 2, pages 827–830 Vol.2, June.

Mahedero, J. P. G., MartÍnez, A., Cano, P., Koppenberger, M., and Gouyon, F. (2005). Natural language processing of lyrics. In *Proceedings of the 13th Annual ACM International Conference on Multimedia*, MULTIMEDIA '05, pages 475–478, New York, NY, USA. ACM.

Mihalcea, R. and Strapparava, C. (2012). Lyrics, music, and emotions. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 590–599, Jeju Island, Korea, July. Association for Computational Linguistics.

Mohammad, S., Bravo-Marquez, F., Salameh, M., and Kiritchenko, S. (2018). Semeval-2018 task 1: Affect in tweets. In *Proceedings of the 12th international workshop on semantic evaluation*, pages 1–17.

Mohammad, S. (2018). Obtaining reliable human ratings of valence, arousal, and dominance for 20,000 english words. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 174–184.

Parisi, L., Francia, S., Olivastri, S., and Tavella, M. S. (2019). Exploiting synchronized lyrics and vocal features for music emotion detection. *CoRR*, abs/1901.04831.

Pauwels, J. and Sandler, M. (2019). A web-based system for suggesting new practice material to music learners based on chord content. In *Joint Proc. 24th ACM IUI Workshops (IUI2019)*.

Pauwels, J., Xambó, A., Roma, G., Barthet, M., and Fazekas, G. (2018). Exploring real-time visualisations to support chord learning with a large music collection. In *Proc. 4th Web Audio Conf. (WAC 2018)*.

Russell, J. A. (1980). A circumplex model of affect. *Journal of personality and social psychology*, 39(6):1161.

Sterckx, L. (2014). *Topic detection in a million songs*. Ph.D. thesis, PhD thesis, Ghent University.

Stöter, F.-R., Uhlich, S., Liutkus, A., and Mitsufuji, Y. (2019). Open-unmix-a reference implementation for music source separation. *Journal of Open Source Software*.

Tagg, P. (1982). Analysing popular music: theory, method and practice. *Popular Music*, 2:37–67.

Vanni, L., Ducoffe, M., Aguilar, C., Precioso, F., and Mayaffre, D. (2018). Textual deconvolution saliency (tds): a deep tool box for linguistic analysis. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 548–557.

Warriner, A. B., Kuperman, V., and Brysbaert, M. (2013). Norms of valence, arousal, and dominance for 13,915 english lemmas. *Behavior research methods*, 45(4):1191–1207.

Watanabe, K., Matsubayashi, Y., Orita, N., Okazaki, N., Inui, K., Fukayama, S., Nakano, T., Smith, J., and Goto, M. (2016). Modeling discourse segments in lyrics using repeated patterns. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1959–1969.

Xia, Y., Wang, L., Wong, K.-F., and Xu, M. (2008). Sentiment vector space model for lyric-based song sentiment classification. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies: Short Papers*, HLT-Short '08, pages 133–136, Stroudsburg, PA, USA. Association for Computational Linguistics.

Yang, D. and Lee, W. (2009). Music emotion identification from lyrics. In *2009 11th IEEE International Symposium on Multimedia*, pages 624–629, Dec.

12. Language Resource References

Bertin-Mahieux, T., Ellis, D. P., Whitman, B., and Lamere, P. (2011). The million song dataset. In *Proceedings of the 12th International Conference on Music Information Retrieval (ISMIR 2011)*.

Meseguer-Brocal, G., Peeters, G., Pellerin, G., Buffa, M., Cabrio, E., Faron Zucker, C., Giboin, A., Mirbel, I., Hennequin, R., Moussallam, M., Piccoli, F., and Fillon, T. (2017). WASABI: a Two Million Song Database Project with Audio and Cultural Metadata plus WebAudio enhanced Client Applications. In *Web Audio Conference 2017 – Collaborative Audio #WAC2017*, London, United Kingdom, August. Queen Mary University of London.